Assignment 7: Written Exercises II

by the Staff of CSE 415

Due March 10, 2021 via GradeScope

This is an individual-work assignment. The staff will not tolerate collaboration, even if some cases were overlooked in the past.

Prepare your answers in a neat, easy-to-read PDF. Our grading rubric will be set up such that when a question is not easily readable or not correctly tagged or with pages repeated or out of order, then points will be be deducted. However, if all answers are clearly presented, in proper order, and tagged correctly when submitted to Gradescope, we will award a 5-point bonus.

If you choose to choose to typeset your answers in Latex using the template file for this document, please put your answers in blue while leaving the original text black.

Paul G. Allen School of Computer Science and Engineering, University of Washington

1 Joint Distributions and Factoring

(35 points) Consider the joint probability distribution below.

A	B	C	$P(A, B, C)$
true	true	true	0.040
true	true	false	0.010
true	false	true	0.360
true	false	false	0.090
false	true	true	0.035
false	true	false	0.015
false	false	true	0.315
false	false	false	0.135

(a) (12 points) For each of the three pairs of random variables (A, B), (A, C), and (B, C), provide a computation to prove that either (i) the two variables are independent, or (ii) they are dependent. If you provide any marginal or other derived distributions, make sure they are clearly identified, e.g., "$P(B|A, C)$:", etc.

$P(A = True) = P(A = True|B = True, C = True) + P(A = True|B = True, C = False) + P(A = True|B = False, C = True) + P(A = True|B = False, C = False) = 0.04 + 0.01 + 0.36 + 0.09 = 0.50$
Similarly,

$$P(A = True) = 0.5, P(A = False) = 0.5$$

$$P(B = True) = 0.1, P(B = False) = 0.9$$

$$P(C = True) = 0.75, P(C = False) = 0.25$$

To prove independence between A and B, we can prove that $P(A, B) = P(A)P(B)$ for all values of A and B. If it is not, A and B are dependent.

$P(A = True, B = True) = P(A = True, B = True|C = True) + P(A = True, B = True|C = False) = 0.05 = P(A = True)P(B = True)$

$P(A = True, B = False) = P(A = True, B = False|C = True) + P(A = True, B = False|C = False) = 0.45 = P(A = True)P(B = False)$

$P(A = False, B = True) = P(A = False, B = True|C = True) + P(A = False, B = True|C = False) = 0.05 = P(A = False)P(B = True)$

$P(A = False, B = False) = P(A = False, B = False|C = True) + P(A = False, B = False|C = False) = 0.45 = P(A = False)P(B = False)$

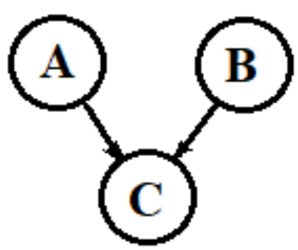
Therefor, A and B are independent. To save space, I will not list all the computations for the rest of pairs, but with the similar methods, we can get that A and C are dependent because

$$P(A = True, C = True) = 0.4 \neq P(A = True)P(C = true) = 0.375$$

and B and C are independent.

Therefore, A and B are independent, B and C are independent, but A and C are dependent.

(b) (10 points) Suppose that somebody has suggested that the joint distribution can be factored in a way that corresponds to the graph below. Prove either that (i) it cannot be factored according to this structure (i.e., some conditional independence assumption would be violated), or (ii) it can be factored according to this structure (i.e., provide the factorization with marginal distributions for A and B, and a CPT (conditional probability table) for $P(C|A, B)$.



It cannot be factored according to this structure. Because according to this structure, C is conditional dependent on A and B, but in fact, C is only conditional dependent on A and it is independent with B.

(c) (3 point) Whether or not our joint distribution can actually be factored according to this two-parent, one-child structure, explain how many "free parameters" would be involved in the specification.

No, our joint distribution can not actually be factored according to this two-parent, one-child structure because C only has one parent, which is A. There are 2 values of A whose probabilities must add up to 1, so one of it will be free parameter, and the other one will be forced to be 1 - the first one. Similarly, B also has one free parameter. For C, which is conditional dependent on A and B, there are also two values whose probabilities will add up to 1. For example, $P(C = T|A = T, B = T) + P(C = F|A = T, B = T) = 1$. However, there are totally four groups of them, thus there are $1 * 4 = 4$ free parameters in total. So, there will be $1 + 1 + 4 = 6$ "free parameters" involved in this specification.

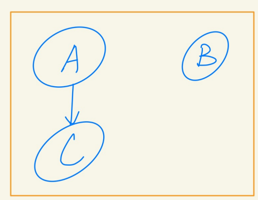
(d) (8 points) Find a factored form for the joint distribution above, using a different graph structure, than the one suggested above, so as to minimize the number of free parameters needed to specify the distribution. Show the needed marginal distributions and conditional probability tables, indicating the numbers of free parameters in each.

1 free

A	P(A)
T	0.5
F	0.5

1 free

B	P(B)
T	0.1
F	0.9



A	C	P(C|A)
T	T	0.8
T	F	0.2
F	T	0.7
F	F	0.3

2 free

C	P(C)
T	0.75
F	0.25

A	C	P(A,C)
T	T	0.40
T	F	0.10
F	T	0.35
F	F	0.15

(e) (2 points) How many free parameters are saved by using this last representation of the joint distribution vs using free parameters of the full joint distribution table at the beginning of this problem? (Be sure to get the correct number of free parameters in the original table before subtracting the number of free parameters of your part d solution.)

Using free parameters of the full joint distribution table requires $8-1 = 7$ free parameters, while only $1+1+2 = 4$ free parameters are needed in the last representation of the joint distribution. So $7 - 4 = 3$ free parameters will be saved.

2 Bayes Nets: D-Separation

(35 points) Consider the Bayes Net graph at the bottom of the page, which represents the topology of a web-server security model. Here the random variables have the following interpretations:

V = Vulnerability exists in web-server code or configs.

C = Complexity to access the server is high. (Passwords, 2-factor auth., etc.)

S = Server accessibility is high. (Firewall settings, and configs are permissive).

A = Attacker is active.

L = Logging infrastructure is state-of-the-art.

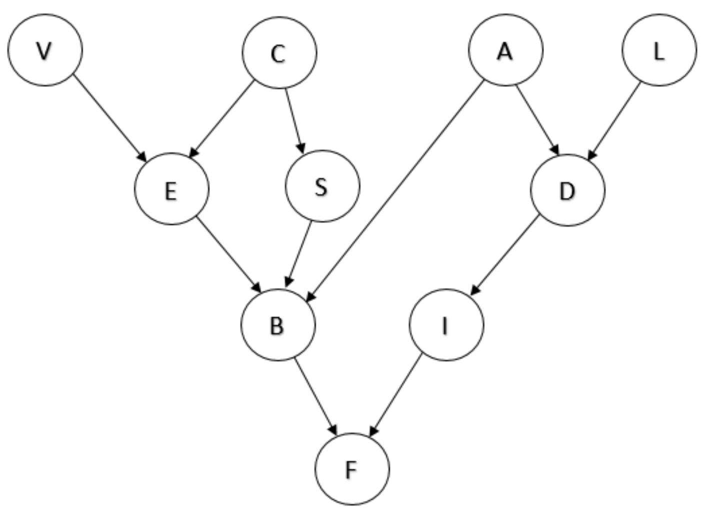
E = Exposure to vulnerability is high.

D = Detection of intrusion attempt.

B = Break-in; the web server is compromised.

I = Incident response is effective.

F = Financial losses are high (due to data loss, customer dissatisfaction, etc.).



For each of the following statements, indicate whether (True) or not (False) the topology of the net guarantees that the statement is true. If False, identify a path ("undirected") through which influence propagates between the two random variables being considered. (Be sure that the path follows the D-Separation rules covered in lecture.) The first one is done for you. (4 points each, except the first one).

(a) $E \perp\!\!\!\perp S$: False (ECS)

(b) $L \perp\!\!\!\perp E \mid F, I$: False(LDIFBE)

(c) $I \perp\!\!\!\perp E \mid A, C$ True

(d) $L \perp\!\!\!\perp C \mid B, F, S$ False(LDIFBEC)

(e) $V \perp\!\!\!\perp L \mid B, D, E$ False(VECSBADL)

(f) $D \perp\!\!\!\perp C \mid E, F, S, V$ True

(g) (15 points) Suppose that the company hired an outside expert to examine the system and she determines that B and E are true: The web server is compromised, and exposure to vulnerability is high. Given this information, your job is to explain to management why getting additional information about D (whether we can detect an intrusion attempt) could have an impact on the probability of V (regarding the existence or non-existence of vulnerabilities). Give your explanation, for the manager of the company, using about between 5 and 15 lines of text, which should be based on what you know about D-separation, applied to this situation. However, your explanation should not use the terminology of D-separation but be in plain English. (You can certainly use words like "influence", "probability", "given", but not "active path", "triple", or even "conditionally independent").

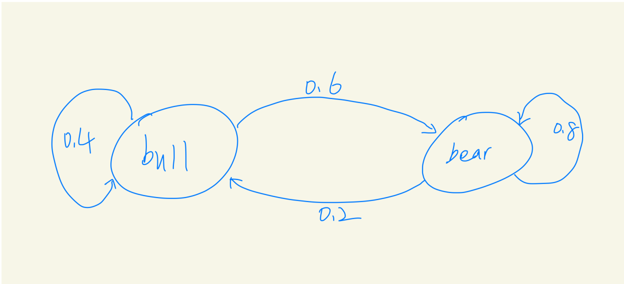
Both S (server accessibility is high) and A (attacker is active) can lead to the appearance of B (logging infrastructure is state-of-art), which is already true. And A can lead to the observation of D (detection of intrusion at tact). So if we detect D, the probability of A will increase, which will make S less likely to be observed. Then the probability of C being observed will decrease given the fact that S is dependent on C. In addition, provided that E is true, either V or C or both are observed. Since the probability of C decreases, the probability of V being true will increase. Therefore, the detection of D could have increased the probability of V.

3 Markov Models

(30 points) According to an unnamed source, the stock market can be modeled using a Markov model, where there are two states "bull" and "bear." The dynamics of the model are given in the table below:

S_{t-1}	S_t	$P(S_t \mid S_{t-1})$
bull	bull	0.4
bull	bear	0.6
bear	bull	0.2
bear	bear	0.8

(a) (6 points) Draw a visual state-transition diagram to represent the conditional probability table of this Markov model.



(b) (10 points) Suppose it's given that S_0 = bull. Perform 3 rounds of the mini-forward algorithm and calculate $P(S_1), P(S_2), P(S_3)$ for each outcome of *bull* and *bear*. Show your work.

$P(S_1 = bull) = P(S_1 = bull|S_0 = bull) = 0.4$
$P(S_1 = bear) = P(S_1 = bear|S_0 = bull) = 0.6$

$P(S_2 = bull) = P(S_2 = bull|S_1 = bull)P(S_1 = bull) + P(S_2 = bull|S_1 = bear)P(S_1 = bear) = 0.4 * 0.4 + 0.2 * 0.6 = 0.28$
$P(S_2 = bear) = P(S_2 = bear|S_1 = bull)P(S_1 = bull) + P(S_2 = bear|S_1 = bear)P(S_1 = bear) = 0.6 * 0.4 + 0.8 * 0.6 = 0.72$

$P(S_3 = bull) = P(S_3 = bull|S_2 = bull)P(S_2 = bull) + P(S_3 = bull|S_2 = bear)P(S_2 = bear) = 0.4 * 0.28 + 0.2 * 0.72 = 0.256$

$P(S_3 = bear) = P(S_3 = bear|S_2 = bull)P(S_2 = bull) + P(S_3 = bear|S_2 = bear)P(S_2 = bear) = 0.6 * 0.28 + 0.8 * 0.72 = 0.744$

(c) (10 points) Compute the stationary probabilities for *bull* and *bear*.

$P_{infinity}(bull) = P(bull|bull)P_{infinity}(bull) + P(bull|bear)P_{infinity}(bear)$
$= 0.4P_{infinity}(bull) + 0.2P_{infinity}(bear)$

$P_{infinity}(bear) = P(bear|bull)P_{infinity}(bull) + P(bear|bear)P_{infinity}(bear)$
$= 0.6P_{infinity}(bull) + 0.8P_{infinity}(bear)$

$P_{infinity}(bull) + P_{infinity}(bear) = 1$

Solving these linear equations, we can get that

$$P_{infinity}(bull) = 0.25$$

$$P_{infinity}(bear) = 0.75$$

(d) (4 points) Suppose a startup uses this model to predict the state of the market in the future. Their analysis software can make a precise observation about the state of the market of previous day S_{t-1} as soon as markets open each day t. One night, a power surge destroyed the server containing all the historic market records. The next day, you are called in to assess the damage. How has this event affected the accuracy of the market predictions of this startup? Give an explanation for why you arrived at your assessment.

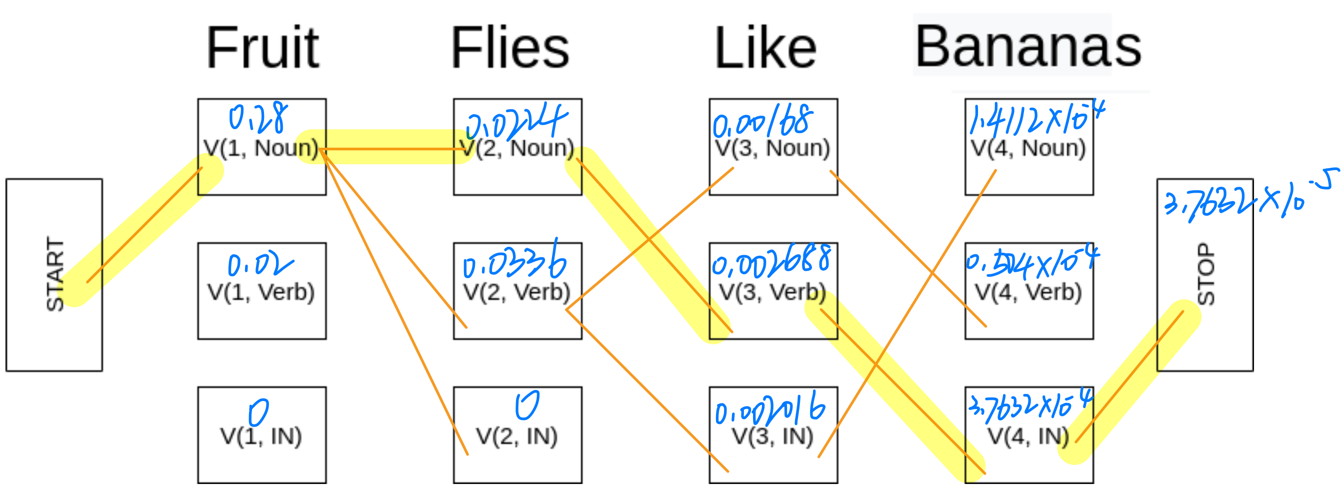
It will have no effect on the accuracy of the market predictions because in Markov model, the probability of an event in day t only depends on the previous day, t - 1. So, it will be unnecessary to store all the historic market records. Since the model will know the performance of (t-1) day's markets as markets open on day t, it will continue making accurate prediction.

4 The Viterbi Algorithm

(35 Points) One of the applicatiions where the Viterbi algorithm can help us is in POS (Part-Of-Speech) tagging in Natural Language Processing. That is, given a sequence of words (e.g., a sentence), assign the correct grammatical tag (Noun, Verb, etc.) to each word based on its definition and context.

- Read more about POS-tagging here
- Read more about the Viterbi algorithm here
- Watch a video to help you understand how to approach this problem here

In this problem we perform POS tagging using the Viterbi algorithm for a small sentence and only considering Nouns (N), Prepositions or subordinating conjunctions (IN) and Verbs (V). While we do not use any other POS tags in this question, if you are curious you can find a list of them here. For the setup below we will be using the Viterbi algorithm to first determine the score for each POS tag for each word and then determine the most probable POS-tag sequence for the entire sentence based on the scores we have calculated.



Let Y_i be random variables representing the emissions (at position i) which are English words and X_i be random variables representing the POS tags. $P(y_i|x_i)$ represents the emission probability of word y_i given tag x_i and $P(x_i|x_{i-1})$ be the probability of transitioning to tag

x_i from tag x_{i-1}. Therefore, for computing score V_{i,x_i} for each state, we have:

$$V_{i,x_i} = \max_{x_1 \dots x_{i-1}} P(x_1 \dots x_i, y_1 \dots y_i) = \max_{x_{i-1}} P(y_i|x_i)P(x_i|x_{i-1})V_{i-1,x_{i-1}}$$

(a) (25 points) In this part we want to compute the maximum score of the tags for each word. Below we have provided the emission and transition probabilities, you will have to use them and the formula above to calculate the score for each state. Show all calculations (you can use a separate page if you want but be sure to include it in the Gradescope submission), please put your maximum score above the corresponding tag box (e.g. put a number above $V(1, N)$ box).

$V(1, N) = P(Fruit|N)P(N|Start)V(0, start) = 0.4 * 0.7 * 1 = 0.28$
$V(1, V) = P(Fruit|V)P(V|Start)V(0, start) = 0.1 * 0.2 * 1 = 0.02$
$V(1, IN) = P(Fruit|IN)P(IN|Start)V(0, start) = 0 * 0.1 * 1 = 0$

$V(2, N) = max[P(Flies|N)P(N|N)V(1, N), P(Flies|N)P(N|V)V(1, V),$
$P(Flies|N)P(N|IN)V(1, IN)] = max[0.2 * 0.4 * 0.28, 0.4 * 0.5 * 0.02, 0.2 * 0.7 * 0] =$
0.0224
$V(2, V) = max[P(Flies|V)P(V|N)V(1, N), P(Flies|V)P(V|V)V(1, V),$
$P(Flies|V)P(V|IN)V(1, IN)] = max[0.4 * 0.3 * 0.28, 0.4 * 0.1 * 0.02, 0.4 * 0.1 * 0] =$
0.0336
$V(2, IN) = max[P(Flies|IN)P(IN|N)V(1, N), P(Flies|IN)P(IN|V)V(1, V),$
$P(Flies|IN)P(IN|IN)V(1, IN)] = max[0 * 0.1 * 0.28, 0 * 0.2 * 0.02, 0 * 0.1 * 0] = 0$

$V(3, N) = max[P(Like|N)P(N|N)V(2, N), P(Like|N)P(N|V)V(2, V),$
$P(Like|N)P(N|IN)V(2, IN)] = max[0.1*0.4*0.0224, 0.1*0.5*0.0336, 0.1*0.7*0] =$
0.00168
$V(3, V) = max[P(Like|V)P(V|N)V(2, N), P(Like|V)P(V|V)V(2, V),$
$P(Like|V)P(V|IN)V(2, IN)] = max[0.4*0.3*0.0224, 0.4*0.1*0.0336, 0.4*0.1*0] =$
0.002688
$V(3, IN) = max[P(Like|IN)P(IN|N)V(2, N), P(Like|IN)P(IN|V)V(2, V),$
$P(Like|IN)P(IN|IN)V(2, IN)] = max[0.3*0.1*0.0224, 0.3*0.2*0.0336, 0.3*0.1*$
$0] = 0.002016$

$V(4,N) = max[P(Bananas|N)P(N|N)V(3,N), P(Bananas|N)P(N|V)V(3,V),$
$P(Like|N)P(N|IN)V(3,IN)] = max[0.1 * 0.4 * 0.00168, 0.1 * 0.5 * 0.002688, 0.1 * 0.7 * 0.002016] = 0.00014112$
$V(4,V) = max[P(Bananas|V)P(V|N)V(3,N), P(Bananas|V)P(V|V)V(3,V),$
$P(Like|V)P(V|IN)V(3,IN)] = max[0.1*0.3*0.00168, 0.1*0.1*0.002688, 0.1*0.1*0.002016] = 0.0000504$
$V(4,IN) = max[P(Bananas)P(IN|N)V(3,N), P(Bananas|IN)P(IN|V)V(3,V),$
$P(Bananas|IN)P(IN|IN)V(3,IN)] = max[0.7*0.1*0.00168, 0.7*0.2*0.002688, 0.7*0.1*0.002016] = 0.00037632$
$V(5, Stop) = max[P(STOP|N)\cdot V(4,N), P(STOP|V)\cdot V(4,V), P(STOP|IN)\cdot \underline{V(4,IN)}] = 3.7632 \times 10^{-5}$

Transition model:

$$\begin{array}{rrrr}
P(N|\text{START}) = 0.7 & P(N|N) = 0.4 & P(N|V) = 0.5 & P(N|IN) = 0.7 \\
P(V|\text{START}) = 0.2 & P(V|N) = 0.3 & P(V|V) = 0.1 & P(V|IN) = 0.1 \\
P(IN|\text{START}) = 0.1 & P(IN|N) = 0.1 & P(IN|V) = 0.2 & P(IN|IN) = 0.1 \\
P(\text{STOP}|\text{START}) = 0.0 & P(\text{STOP}|N) = 0.2 & P(\text{STOP}|V) = 0.2 & P(\text{STOP}|IN) = 0.1
\end{array}$$

Emission model:

$$\begin{array}{rrr}
P(\text{fruit}|N) = 0.4 & P(\text{fruit}|V) = 0.1 & P(\text{fruit}|IN) = 0.0 \\
P(\text{flies}|N) = 0.2 & P(\text{flies}|V) = 0.4 & P(\text{flies}|IN) = 0.0 \\
P(\text{like}|N) = 0.1 & P(\text{like}|V) = 0.4 & P(\text{like}|IN) = 0.3 \\
P(\text{bananas}|N) = 0.1 & P(\text{bananas}|V) = 0.1 & P(\text{bananas}|IN) = 0.7
\end{array}$$

(b) (10 points) Now that we have computed the score, determine the maximum-probability sequence of states by working backwards from the STOP state. Consider the STOP state as the first value of a variable CURRENT-STATE. At each stage, move to the left, to the state that was selected during the arg-maxing for CURRENT-STATE. Then make that state be CURRENT-STATE, and iterate until all the way left at START. Reading the sequence along this path from left to right, you'll have the maximum-probability sequence of states that could give rise to the sentence "Fruit flies like bananas." Clearly draw the path, and also write down the final maximum-probability tag sequence.

5 Disambiguating Syntax with PCFGs

(35 points) Consider the sentence "Caitlin watched the boy with a telescope." This might mean that, for example, Caitlin used a telescope to watch the boy, or rather differently it might mean that Caitlin watched a boy who had a telescope.

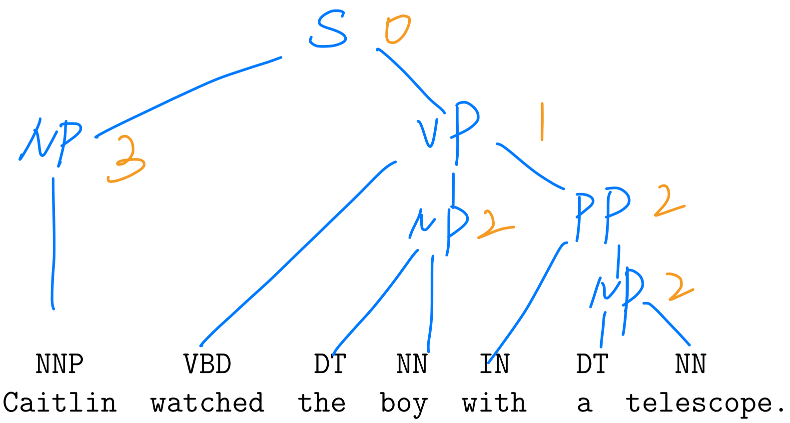
A grammar relevant to this example is given below. Consider the number at the right of a production to be the conditional probability of applying that production given that the symbol to be expanded, during a derivation, is the symbol on the left-hand side of the production.

In this problem, you'll convert probabilities of productions into scores. Then, with the given probabilistic context-free grammar, you will find two legal parses for the sentence, and compute a score for each parse. You'll then convert the overall parse scores back to probabilities of each parse. Then you'll identify the more probable parse using the parse probabilities.

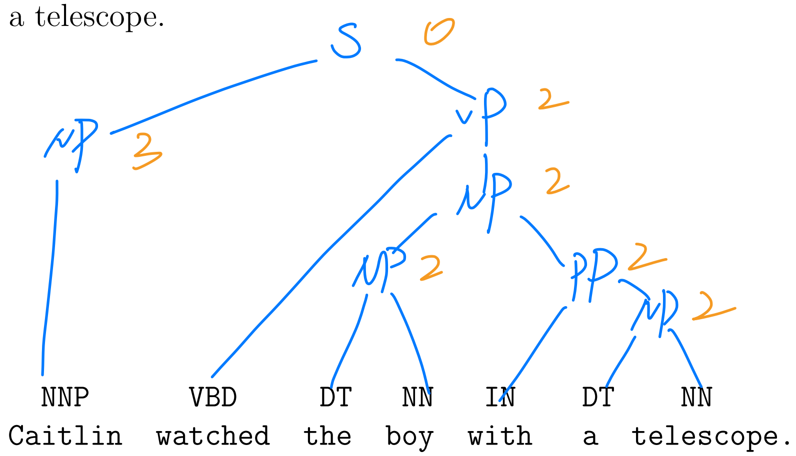
```
S    ::=  NP   VP        1.000    0
NP   ::=  NN             0.250    2
NP   ::=  NP   PP        0.250    2
NP   ::=  NNP            0.125    3
NP   ::=  DT   NN        0.250    2
NN   ::=  NNP            0.125    3
VP   ::=  VBD  NP   PP   0.500    1
VP   ::=  VBD  NP        0.250    2
PP   ::=  IN   NP        0.250    2
```

(a) (8 points) Scores for each production rule: Convert each probability into a score by taking score $= -\log_2(p)$. Write the scores next to the probabilities above.

(b) (6 points) Find parse number 1. The parse will assume that the terminal symbols have been converted to non-terminals as shown. Make this parse correspond to the interpretation that Caitlin has the telescope.



(c) (6 points) Parse number 2. Make this parse according to an interpretation in which the boy has a telescope.



(d) (6 points) Total score 10 and overall probability 2^{-10} for parse number 1.

(e) (6 points) Total score13 and overall probability 2^{-13} for parse number 2.

(f) (3 point) Which parse is more probable? parse number 1

You are welcome but not required to use the online parser at https://parser.kitaev.io/ in this problem. Note that it doesn't return all parses of a sentence.

6 The Laws of Robotics and Ethics in AI

(30 points) Read the short story *Reason* by Isaac Asimov and the article *Artificial Intelligence: Does Consciousness Matter?* Answer the questions listed below the links (5 points each).

Short Story: https://canvas.uw.edu/courses/1431895/files?preview=74233603

Article: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6614488/

(a) Who was Isaac Asimov and why is he well known among computer scientists?

Issac Asimov was a well known American writer who wrote many science fictions, short stories, and popular science books. He was well known among computer scientists because of his three laws of robotics.

(b) What are Asimov's 3 Laws of Robotics?

First Law: A robot may not injure a human or through inaction, allow a human to come to harm.

Second Law: A robot must obey the orders given it by human beings, unless such orders would conflict with the first law.

Third Law: A robot must protect its own existence, as long as such protection does not conflict with the first or second law

(c) Is the ordering of the laws important? If so, explain why. If not, explain why not. (You may find referring to this resource helpful: https://xkcd.com/1613/))

Yes, the ordering of the laws is important because the order represents our expectations of robots. When we create a new robot, first of all, we want to make sure that it will protect and not harm us. Then we expect it to obey our orders to help us. And finally, based on the first two laws, we want the robot to live as long as possible to help us. If we change orders, just as the comic depicts, the world may result in chaos. For example, if we swap the order of the second and the third law, we, expect the robot to protect itself before listening to us. So, the robot may disobey our order that will threaten its existence, such as exploring Mars.

(d) Do you think Cutie's actions are consistent with the three laws (give specific examples that support your answer)?

Yes, Cutie's actions are consistent with the three laws. It seems like Cutie breaks the second law when he disobeys Powell's order, but actually his decision is subject to the first law. For example, Cutie doesn't allow Powell and Donavan to take back the control room because he believes that he is better suited than Powell or Donavan to do the job to prevent the catastrophe. So, as a matter of fact, Cutie obeys all three laws.

(e) Do you think Cutie is 'conscious' or 'sentient'? In your response, please define what you think these terms mean and how they might apply to Cutie.

I think Cutie is conscious and sentient. In my definition, conscious is the capability of self-awareness and sentient is the capability of feeling emotions. Cutie exhibits curiosity as to his existence: he is aware of his existence. And he can introspect himself to come out with a satisfying answer – "I, myself, exist because I think". Besides, Cutie shows emotions or feelings in the story. For example, "There was a pity in Cutie's voice". That's why I think he is conscious and sentient.

(f) In this story, Powell tells Cutie, "If you don't satisfy us, you will be dismantled." If he carried through on his threat, would such an action be ethical? Please refer to the article on consciousness and AI in your response.

It would be ethical if Cutie doesn't develop consciousness and sentiment because, in this case, Cutie is just an object, a machine. No one or few people will condemn a person for dismantling an object. But once Cutie develops consciousness, such actions will be unethical because now people can develop emotional bonds with him, like social robots described in the article. "It is in line with our social values to treat robots more like pets than like mere things", and since it is unethical to abuse animals, it is also unethical to abuse a robot with consciousness.